Security of Spectrum Learning in Cognitive Radios

Behnam Bahrak and Jung-Min "Jerry" Park
Department of Electrical and Computer Engineering
Virginia Tech, Blacksburg, VA, 24061
Email: {bahrak, jungmin}@vt.edu

Abstract

Due to delay and energy constraints, a *cognitive radio* may not be able to perform spectrum sensing in all available channels. Therefore, a sensing policy is needed to decide which channels to sense. The *channel selection* problem is the problem of designing such a sensing policy to maximize throughput while avoiding interference to primary users. The channel selection problem can be formulated as a reinforcement learning problem. Channel selection schemes that employ reinforcement machine learning algorithms are vulnerable to *belief manipulation attacks* that contaminate the knowledge base of the learning algorithms. In this paper, we analyze the security of channel selection algorithms that are based on reinforcement learning and propose mitigation techniques that make these algorithms more robust against belief manipulation attacks.

I. Introduction

It is widely believed that cognitive radios (CRs) are one of the key technologies that can address the spectrum scarcity problem. It is expected that they will play an important role in maximizing spectrum utilization and help satisfy the QoS requirements of a number of important communications applications—from emergency first responders' public safety networks to military tactical networks. CRs often employ software-defined radio platforms that are capable of executing complex computational tasks to communicate efficiently without causing interference to licensed (a.k.a. primary) users. A specialized software module within a CR called the *cognitive engine* performs the aforementioned tasks, such as the optimization of the transmission/reception

(TX/RX) parameters and execution of spectrum sensing and spectrum access strategies. Most of the tasks performed by a cognitive engine require the use of machine learning algorithms, especially if those tasks need to be carried out in a distributed manner.

Considering the computing limitations and energy constraints of a battery-powered CR, a CR may not be able to perform full-spectrum sensing (i.e., sense all available spectrum bands) because of its prohibitive cost. Therefore, a spectrum sensing policy at the medium access control (MAC) layer is needed to decide which set of channels to sense. The *channel selection problem* is the problem of designing such a sensing policy. The optimal channel selection strategy for an unlicensed user (i.e., secondary user) is based on the availability statistics of the channels. The availability of the channels is determined by the presence/absence of primary user signals in those channels. The channels' availability statistics are initially unknown to a secondary user and need to be estimated using sensing samples. The critical tradeoff that the cognitive engine faces in each timeslot is between transmission ("exploitation") on the channel that has the highest expected reward (e.g., throughput) and channel sensing ("exploration") to get more information about the expected rewards of the other channels. The exploitation vs. exploration tradeoff problem, such as the one just described, is central to an area of machine learning known as *reinforcement learning*.

Spectrum learning is the process of learning the spectrum statistics (i.e., primary user occupancy information), which is crucial to enable CRs to sense/interpret their spectrum environment and make intelligent decisions to achieve efficient communication. Although spectrum learning is beneficial for CRs, it can pose a serious security vulnerability. A radio that can learn has the potential to be taught by malicious entities in an adversarial environment. This kind of threat may have a long-lasting impact on the cognitive radio network.

As mentioned above, the design of the optimal sensing policy can be formulated as a *reinforcement learning* (RL) problem. When the channels are assumed to be independent, it can be formulated as a special class of RL problems known as a *restless multi-armed bandit* process. Recent results (i.e., [1] – [4]) show that a surprisingly simple *myopic policy* that ignores the impact of the current action on the future reward is optimal when channels are identical. In this paper, we show that in adversarial environments, where an active attacker performs belief manipulation attacks against the machine learning algorithms executed on a cognitive engine, the myopic policy is no longer optimal and a *softmax policy* that exploits some level of randomness

outperforms the myopic policy.

Our contributions can be summarized as follows:

1) We derive closed-form expressions for the throughput of cognitive radios in an adversarial environment for the two-channel case in two channel selection policies, viz myopic policy and softmax policy.
2) We derive the attacker's optimal attack strategy and the cognitive engine's optimal defense strategy by solving respective optimization problems for more than two non-identical channels.
3) We identify and discuss two fundamental trade-offs in the security of spectrum learning: (1) the attackers tradeoff between the attack probability and the number of required observations for attack detection (i.e., the time that the attack detection system needs to detect an attack); and (2) the channel selection systems tradeoff between attack resilience and performance of a given channel selection policy.
4) We prove that for sufficiently large attack probabilities, a softmax policy with a proper choice of parameters outperforms the myopic policy for all possible attacker's strategies.

The rest of this paper is organized as follows. In Section II, the work related to this paper are discussed. Section III provides the channel selection system model and the attack model. In Section IV, we analyze sensing policies in an adversarial environment for a cognitive radio system with two channels and Section V extends the result to a cognitive radio system with more than two channels. Finally Section VI concludes the paper.

II. Related Work

The formulation of spectrum learning problem as a restless multi-armed bandit process is investigated in [1]– [6]. It is proved that when channels are identical and independent the myopic policy is the optimal policy [2]– [4], and that this policy is a special case of Whittle's index policy for the restless bandit problem which can be computed for non-identical channels as well [5]. An asymptotically optimal policy is proposed in [6] for a more realistic case where the policy does not require any prior statistical knowledge about the traffic pattern and the channels are different. Despite all these work that assume a non-adversarial environment, this paper investigates the spectrum learning problem in an adversarial environment where active attackers aim to reduce the throughput of the cognitive radio network.

The security of machine learning algorithms that have been applied to applications such as intrusion detection systems (IDS) and spam filters is investigated thoroughly in [7]- [11]. In these papers, the authors discuss how an adversary can maliciously mistrain a learning system in an IDS and how an attacker may contaminate the knowledge base of a spam email filtering system to bypass the filtering. In [8] and [10] different kinds of attacks against machine learning algorithms are introduced and a variety of potential defenses against those attacks are proposed.

In the context of cognitive radios, security of machine learning algorithms that are used for signal classification are addressed in [18] and [19], but the types of learning algorithms that are analyzed is different from the algorithms in this paper. To the best of our knowledge, this paper describes the first analysis of a reinforcement learning algorithm's vulnerability against belief manipulation attacks to cognitive radios.

III. The Channel Selection System and the Attack Model

In this section, we introduce the channel selection system and establish the attack model.

A. Channel Selection System Model

We consider a general dynamic spectrum access system where a user has access to N independent and stochastically non-identical parallel Gillbert-Elliot channels [12], and chooses one channel to sense and access in each time slot, aiming to maximize its expected long-term reward (i.e., throughput).

As illustrated in Figure 1, the state of the k-th channel—either idle (1) or busy (0)—indicates that the channel is unused by primary users or it is occupied. The transitions between these two states follow a Markov chain with transition probabilities $\{p_{ij}^k\}_{i,j=0,1}$. We assume that these transition probabilities for all channels are learnt in a non-adversarial environment before the system starts operating and thus the transition probabilities are known to the system. We also assume that $p_{11}^k > p_{01}^k$ or equivalently the channel states in two consecutive time slots are positively correlated. Note that this assumption is only used for derivation of closed-form expressions and can easily be relaxed by separately considering the case where this assumption does not hold. Due to its limited sensing and access capability, a secondary user chooses one of the N channels to sense and access in each slot. Designing an optimal sensing policy that governs the channel selection at each time slot can be formulated as a restless multi-armed

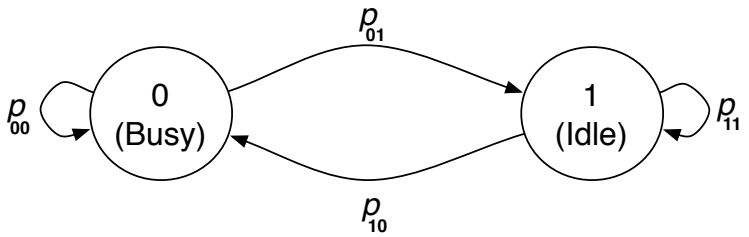


Fig. 1. A Gillbert-Elliot channel.

bandit process for independent channels. We denote $S_k(t)$ as the state of channel k in slot t that is given by the two-state Markov chain in Figure 1. Let $S(t) = [S_1(t), \cdots, S_N(t)] \in \{0,1\}^N$ denote the full system state.

The channel selection system is reward-based. At each time slot the secondary user selects one of the N channels to sense. If the sensed channel is occupied by primary user signals, the user collects no reward; otherwise it accesses the channel and collects one unit of reward. The system keeps periodical sensing and transmitting on that channel until a primary user appears on the channel or a jamming attack prevents the user from transmission on the channel. The secondary user's aim is to maximize the *throughput* (reward) over a horizon of T slots by choosing an optimal sensing policy.

Due to limited sensing (i.e., sensing only one channel out of N channels), the full system state ($S(t)$) in slot t is not observable. However, it has been shown that a sufficient statistic for optimal decision making is given by the conditional probability that each channel is in state 1, given all past observations and decisions [13]. Referred to as the belief vector, we denote this sufficient statistic by $\Omega(t) = [\omega_1, \cdots, \omega_N]$, where $\omega_k(t)$ is called the belief value of channel k which is equivalent to the conditional probability that $S_k(t) = 1$ given all past observations and decisions for that channel. Given the sensing action $a(t) = k$ (the channel that is selected to be sensed in slot t) and the observation $S_k(t)$ in slot t, the belief vector for slot $t+1$ can be updated via Bayes rule through the following equation:

$$\omega_k(t+1) = \begin{cases} p_{11}^k, & a(t) = k, S_k(t) = 1 \\ p_{01}^k, & a(t) = k, S_k(t) = 0 \\ \Gamma(\omega_k(t)), & a(t) \neq k \end{cases}, \tag{1}$$

where $\Gamma(x) = xp_{11}^k + (1-x)p_{01}^k$.

A sensing policy π specifies a sequence of functions $\pi = [\pi_1, \cdots, \pi_t]$, where π_t maps the belief vector $\Omega(t)$ to a sensing action $a(t)$. Multi-channel opportunistic access can thus be formulated as the following stochastic optimization problem:

$$\pi^* = \arg\max_{\pi} E_\pi[\sum_{t=1}^{T} R_{\pi_t(\Omega(t))}(t)|\Omega(1)], \tag{2}$$

where $\pi_t(\Omega(t))$ is the channel selected for sensing and $R_{\pi_t(\Omega(t))}(t)$ is the reward when the belief vector is $\Omega(t)$ and the action $\pi_t(\Omega(t))$ is taken, and $\Omega(1)$ is the initial belief vector. If no information about the initial system state is available, each entry of $\Omega(1)$ can be set to the stationary distribution ω_0^k of the underlying Markov chain:

$$\omega_0^k = \frac{p_{01}^k}{p_{01}^k + p_{10}^k}. \tag{3}$$

Let $V_t(\Omega(t))$ be the value function which represents the maximum expected total reward that can be obtained starting from slot t given the current belief vector $\Omega(t)$. Given that the user selects channel k and observes $S_k(t)$ in slot t, the maximum expected reward consists of the following two parts:

1) The expected immediate reward:

$$E[R_k(t)] = E[S_k(t)] = \omega_k(t).$$

2) The maximum expected future reward:

$$V_{t+1}(\tau(\Omega(t)|k, S_k(t))),$$

where $\tau(\Omega(t)|k, S_k(t))$ denotes the updated belief vector for slot $t+1$ as given in (1). If we maximize over all channel selections, we obtain the following optimization equation:

$$V_t(\Omega(t)) = \max_{k=1,\cdots,N} \{\omega_k(t) + V_{t+1}(\tau(\Omega(t)|k, S_k(t)))\}.$$

Because the value function is limited to horizon T, we have: $V_T(\Omega(T)) = \max_{k=1,\cdots,N} \omega_k(T)$ and $V_t(\Omega(t)) = 0$ for $t > T$. Theoretically, the optimal policy π^* and its performance $V_1(\Omega(1))$ can be obtained by solving the above dynamic programming problem. However, because of the

impact of the current action on the future reward and the uncountable space of the belief vector, obtaining the optimal solution via the above recursive equations is computationally prohibitive.

B. Attack Model

We assume that there exists a single attacker in the environment who tries to decrease the throughput of secondary users by preventing them from transmission in some time slots, by employing adaptive interference techniques (or cognitive jamming attacks). Although many defenses against these attacks have been proposed, but none of these defenses is perfect or can address all classes of attackers. While such defenses can restrict the instantaneous effect of these attacks, they can not reduce the long-term effect of such attacks on cognitive radio network, when the network is using a learning system as part of its cognitive engine. These attacks gradually contaminate the knowledge base of a cognitive radio and lead the learning system to make wrong decisions which results in degrading the performance of the radio.

The attacker uses the same equipment as secondary users, i.e., it is equipped with a CR that is comparable to a typical secondary user's CR in terms of battery capacity, transmission power, computing power, memory capacity, etc. Therefore the attacker is power-limited and wish to avoid jamming continuously, which quickly drains power and causes fast detection by the attack detection module of cognitive engine. We also assume that the attacker can attack only one channel at each time slot. Suppose that the attacker perform attacks in t_a slots out of T consecutive time slots on average. We denote $\alpha = \frac{t_a}{T}$ as the *attack probability* in the environment. The attacker controls the attack probability in order to cause maximal damage to the network in terms of throughput reduction. It can easily be seen that a higher attack probability will result in a lower throughput for the cognitive radio system.

We assume that the adversary possesses full knowledge about the network and its parameters. We introduce the notion of the attacker's optimal strategy using the following definition:

Definition 1: The attacker's α**-optimal strategy** is a strategy that minimizes the throughput of the target cognitive radio while keeping the attack probability fixed to value α.

C. Attack Detection Model

The cognitive radio network employs a mechanism for monitoring network status and detecting potential malicious activity. This monitoring mechanism is proposed in [20] to protect the network

against jamming attacks. The monitoring can be done by specific monitor nodes in a distributed network and a detection algorithm is employed by the detection module at a monitor node; it takes as input observation samples obtained by the monitor node (i.e., failed transmission/successful transmission) and decides whether there is an attack or not. On one hand the observation window should be small enough, such that the attack is detected in a timely manner and appropriate countermeasures are initiated. On the other hand, this window should be sufficiently large such that the chance of a false alarm or a mis-detection is reduced.

The sequential nature of observations at consecutive time slots motivates the use of sequential detection techniques. A sequential decision rule is efficient if it can provide reliable decisions as fast as possible. There exists a trade-off between detection delay and detection accuracy in a detection scheme, i.e. a faster decision unavoidably leads to higher values of the probability of false alarm P_{FA} and probability of mis-detection P_M while lower values of these probabilities are attained at the expense of detection delay. For given values of P_{FA} and P_M, the detection test that minimizes the average number of required observations (and thus average delay) to reach a decision among all sequential and non-sequential tests is *Wald's Sequential Probability Ratio Test (SPRT)* [15].

In our case, the test is between hypotheses H_0 and H_1 with Bernoulli probability mass functions (p.m.fs) f_0 and f_1. Assume that Y is a random variable with the Bernoulli distribution, where $Y = 1$ denotes a failed transmission event in a slot. H_0 denotes the hypothesis that assumes the absence of an attack, and because the probability of a failed transmission in the absence of attack is p_{10} (i.e. $Pr\{Y = 1\} = p_{10}$), the corresponding p.m.f f_0 has a Bernoulli distribution with parameter $\theta_0 = p_{10}$. Similarly, because the probability of a failed transmission in the presence of an attack with attack probability α is $1 - p_{11}(1 - \alpha)$ (i.e. $Pr\{Y = 1\} = 1 - p_{11}(1 - \alpha)$), H_1 that denotes the hypothesis that assumes the existence of an attack has a Bernoulli p.m.f f_1 with parameter $\theta_1 = 1 - p_{11}(1 - \alpha)$.

The logarithm of likelihood ratio at stage k with observations $x_1, \cdots, x_k$ is:

$$S_k = \sum_{i=1}^{k} \ln \frac{f_1(x_i)}{f_0(x_i)}.$$

where $x_i = 1$, if the system observes a failed transmission at time slot i, and $x_i = 0$ otherwise.

The decision variable is defined as follows:

$$S_k \geq a \Rightarrow \text{ accept } H_1$$

$$S_k < b \Rightarrow \text{ accept } H_0$$

$$b \leq S_k < a \Rightarrow \text{ take another observation}$$

The analysis in [20] shows that using this SPRT, the average number of samples needed for detecting an attack is:

$$E[N|H_1] = \frac{C}{\theta_1 \ln(\frac{\theta_1}{\theta_0}) + (1-\theta_1)\ln(\frac{1-\theta_1}{1-\theta_0})},$$

where C is a fixed positive number. Using this equation for $E[N|H_1]$ we have:

$$\begin{aligned}\frac{\partial E[N|H_1]}{\partial \alpha} &= \frac{\partial E[N|H_1]}{\partial \theta_1} \times \frac{\partial \theta_1}{\partial \alpha} \\ &= \frac{-C(\ln(\frac{\theta_1}{\theta_0}) + \ln(\frac{1-\theta_0}{1-\theta_1}))}{(\theta_1 \ln(\frac{\theta_1}{\theta_0}) + (1-\theta_1)\ln(\frac{1-\theta_1}{1-\theta_0}))^2} \times p_{11}.\end{aligned}$$

Because $\theta_1 > \theta_0$, we have $\frac{\partial E[N|H_1]}{\partial \alpha} < 0$ and consequently $E[N|H_1]$ is a decreasing function of α, i.e., increasing the attack probability would decrease the average number of required observations for attack detection. This poses a fundamental trade-off problem to an attacker: Increasing the attack probability, α, increases the impact of the attack on the target (i.e., lower its throughput), but it also enables a detection module to detect the attack sooner. We define the attacker's cost as the inverse of $E[N|H_1]$:

$$\frac{\theta_1 \ln(\frac{\theta_1}{\theta_0}) + (1-\theta_1)\ln(\frac{1-\theta_1}{1-\theta_0})}{C},$$

Also in order to normalize the attacker's cost, we assume that the constant C is equal to the maximum value of the statement $\theta_1 \ln(\frac{\theta_1}{\theta_0}) + (1-\theta_1)\ln(\frac{1-\theta_1}{1-\theta_0})$ that happens at $\alpha = 1$, i.e. $C = \ln(\frac{1}{p_{10}})$, therefore:

$$\text{Attacker Cost} = \frac{\theta_1 \ln(\frac{\theta_1}{\theta_0}) + (1-\theta_1)\ln(\frac{1-\theta_1}{1-\theta_0})}{\ln(\frac{1}{p_{10}})}. \tag{4}$$

This definition for the attacker's cost shows that by risking detection of the attack by a detection module, the attacker's cost increases. The equation 4 would be used as a measure for the

attacker's cost in the rest of this paper.

IV. Sensing Policies Analysis in an Adversarial Environment with two channels

In this section, we analyze and compare the myopic policy [2] and the softmax policy [14] in a hostile environment for $N = 2$ identical channels, i.e. $p_{ij}^1 = p_{ij}^2 = p_{ij}$.

A. Analysis of the Myopic Policy

The myopic policy explained in this section is identical to the one presented in [2], but in this section, we analyze the performance of this policy in an adversarial environment. A myopic policy ignores the impact of the current action on the future reward and only focuses on maximizing the immediate reward. At any given time slot t, the myopic policy for selecting a channel for sensing can be expressed as follows:

$$\pi^m(t) = \arg \max_{i=1,\cdots,N} \omega_i(t).$$

In [3], the authors proved that for the channel selection system that was introduced in Section 2.1, the myopic policy is the optimal policy for all N. They also showed that the myopic policy has a simple structure that does not require the knowledge of the transition probabilities p_{ij} or updates to the belief vector.

We define the steady-state throughput of the myopic policy as:

$$U^m = \lim_{T\to\infty} \frac{V_{1:T}^m(\Omega(1))}{T},$$

where $V_{1:T}^m(\Omega(1))$ is the expected total reward obtained in T slots under the myopic policy when the initial belief vector is $\Omega(1)$. The key to computing the throughput U is to first find how long a user stays in the same channel. Let us introduce the concept of a *transmission period* (TP), which represents the time that a user stays in the same channel. Let L_k denote the k-th TP. In [3], Zhao et al. showed that under the condition $p_{11} > p_{01}$, the steady-state throughput is:

$$U = 1 - \frac{1}{\overline{L}}, \tag{5}$$

where $\overline{L} = \lim_{K\to\infty} \frac{\sum_{k=1}^{K} L_k}{K}$ denotes the average length of a TP.

Throughput analysis is thus reduced to analyzing the average TP length $\overline{L}$. For $N = 2$, we can derive a closed-form expression of $\overline{L}$ as a function of the attack probability α, which leads to a closed-form expression of the myopic policy throughput $U^m(\alpha)$.

Theorem 1: For $N = 2$, the average TP length of a myopic policy as a function of the attack probability, α, is given by:

$$L^m(\alpha) = 1 + \frac{\overline{\omega}}{1 - p_{11}(1-\alpha)}, \tag{6}$$

where

$$\overline{\omega} = \frac{(1-\alpha)p_{01}^{(2)}}{(1-\alpha)p_{01}^{(2)} - A}, \tag{7}$$

and

$$A = \omega_0(1-\alpha)[1 - \frac{(p_{11}-p_{01})^3(1-p_{11}(1-\alpha))}{1-p_{11}(1-\alpha)(p_{11}-p_{01})}], \tag{8}$$

and

$$p_{01}^{(2)} = \frac{p_{01} - p_{01}(p_{11}-p_{01})^2}{p_{01}+p_{10}}.$$

Proof: From the structure of the myopic policy, $\{L_k\}_{k=1}^{\infty}$ forms a first-order Markov chain for N = 2. When the system is running in an adversarial environment with attack probability α, the transition probabilities of $\{L_k\}_{k=1}^{\infty}$ are given by

$$r_{ij} = \begin{cases} 1 - p_{01}^{(i+1)}(1-\alpha) & j = 1 \\ p_{01}^{(i+1)}(1-\alpha)^{j-1}p_{11}^{j-2}(1-p_{11}(1-\alpha)) & j \geq 2 \end{cases},$$

where $P_{01}^{(j)}$ is the j-step transition probability which is equal to $\omega_0 - \omega_0(p_{11}-p_{01})^j$. Let $\mathbf{R} = \{r_{ij}\}$ denote the transition matrix of $\{L_k\}_{k=1}^{\infty}$ and let $\mathbf{R}(:,k)$ denote the k-th column of $\mathbf{R}$. We have

$$\mathbf{1} - \mathbf{R}(:,1) = \frac{\mathbf{R}(:,2)}{1-p_{11}(1-\alpha)}, \tag{9}$$

and

$$\mathbf{R}(:,k) = \mathbf{R}(:,2)(p_{11}(1-\alpha))^{k-2} \text{ for } k \geq 2, \tag{10}$$

where $\mathbf{1}$ is the unit column vector $[1,1,\cdots]^T$. We denote $\Lambda = [\lambda_1, \lambda_2, \cdots]$ as the stationary

distribution of $\{L_k\}_{k=1}^{\infty}$, i.e. $\Lambda\mathbf{R} = \Lambda$. Thus we have:

$$[\lambda_1, \lambda_2, \cdots]\mathbf{R}(:, k) = \lambda_k. \tag{11}$$

Combining (9), (10) and (11) results in

$$\lambda_1 = 1 - \frac{\lambda_2}{1 - p_{11(1-\alpha)}},\ \lambda_k = \lambda_2(p_{11}(1-\alpha))^{k-2}. \tag{12}$$

Substituting (12) into (11) and solving for λ_2, we get $\lambda_2 = \overline{\omega}(1 - p_{11}(1-\alpha))$, where $\overline{\omega}$ is given in (7). From (12), we can find the stationary distribution as

$$\lambda_k = \begin{cases} 1 - \overline{\omega}, & k = 1 \\ \overline{\omega}(1 - p_{11}(1-\alpha))(p_{11}(1-\alpha))^{k-2} & k > 1 \end{cases}. \tag{13}$$

Using (13) to compute $L^m(\alpha) = \sum_{k=1}^{\infty} k\lambda_k$ results in (6). ■

Using the results of Theorem 1, we can show that $U^m(\alpha)$ is a decreasing function of α, and thus an attacker can lower the target radio's throughput (which is employing myopic sensing) by increasing the attack probability α. However, as we discussed in Section 2.2, increasing α increases the probability that the attack is detected.

B. Analysis of the Softmax Policy

The *softmax* action selection policies are randomized policies where, at time t, the action a_t is chosen at random by the user according to some probability distribution giving more weight to actions which have performed well in the past. The greedy action is given the highest selection probability, but all the others are ranked and weighted according to their accumulated rewards [14]. The most common softmax action selection method uses a Gibbs or Boltzman distribution for the action selection probabilities. It chooses action a at time slot t, with probability

$$p_a(t) = \frac{e^{\omega_a(t)/\tau}}{\sum_{i=1}^{N} e^{\omega_i(t)/\tau}},$$

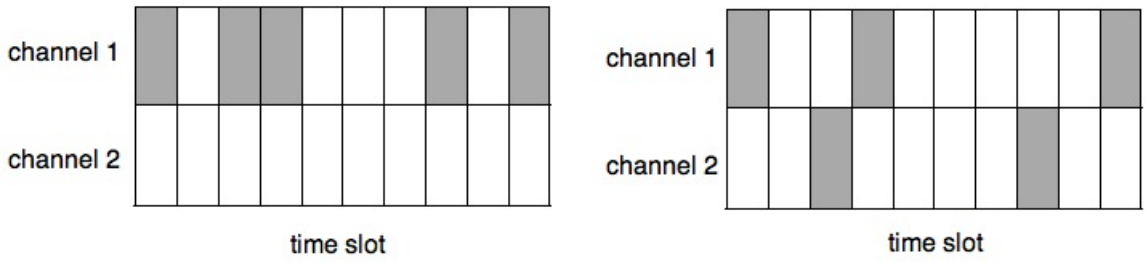
where τ is a positive parameter called the *temperature* and controls the greediness of the policy. High temperatures cause all the actions to be all equiprobable while low temperatures cause a high probability for greedy action, and in the limit as $\tau \to 0$, the softmax policy becomes equivalent to the myopic policy.

In this section, for simplicity and ease of computation, we use a Bernoulli distribution instead of a Boltzmann distribution, i.e., we choose action a at time slot t with probability

$$p_a(t) = \begin{cases} q & \text{if} \quad a = \arg\max_{i=1,2} \omega_i(t) \\ 1-q & \text{if} \quad a = \arg\min_{i=1,2} \omega_i(t) \end{cases} . \tag{14}$$

We define the *main probability* q as the probability of taking greedy action (i.e. selecting the channel that has the highest ω_i). According to the definition of q and the softmax policy, we have $0.5 \leq q \leq 1$ when $N = 2$. Also note that for $q = 1$, the softmax policy reduces to myopic policy, i.e., myopic policy is a special case of softmax policy.

The attacker's optimal strategy for attacking a channel selection system employing the myopic policy is simple: only attack the channel that has the biggest belief value, since the user does not transmit on other channels. The attackers optimal strategy against the softmax policy is not so straightforward.



(a) attacker's strategy with division probability d = 1 (b) attacker's strategy with division probability d = 0.6

Fig. 2. attack strategy examples for fixed attack probability $\alpha = 0.5$

As mentioned in Section 2.2, the α-optimal strategy for an attacker is a strategy that minimizes the throughput of a cognitive radio while keeping the attack probability α fixed. Knowing that the softmax policy uses a fixed main probability q for channel selection (i.e., it uses (14) for channel selection), the attacker divides its attacks between the two channels. We define d as the conditional probability of channel 1 (the greedy option for the policy) being attacked at a given timeslot, assuming that the timeslot is attacked and call it the *division probability*. Figure 2 illustrates the concept of the division probability (the slots colored in gray are the time slots in which an attacker jams a channel). The attacker chooses d such that a cognitive radios throughput is minimized. On the other hand the channel selection system exploits its knowledge about the

optimal strategy of the attacker to select the main probability q such that the throughput is maximized. Assume that $U^s(q,d)$ define the throughput of the radio when the softmax policy uses a main probability of q and the attacker exploits the division probability d, constructing optimal attack strategy and its corresponding optimal channel selection probability distribution can be formulated as the following optimization problems:

Optimization Problem 1:

$$d^* = \min_d U^s(q,d)$$
$$\text{s.t. } 0 \leq d \leq 1$$

Optimization Problem 2:

$$q^* = \max_q U^s(q,d^*)$$
$$\text{s.t. } 0.5 \leq q \leq 1$$

The steady-state throughput of the softmax policy is given by:

$$U^s = \lim_{T\to\infty} \frac{V^s_{1:T}(\Omega(1))}{T},$$

where $V^s_{1:T}(\Omega(1))$ is the expected total reward obtained in T slots under the softmax policy when the initial belief vector is $\Omega(1)$. As shown in Section 3.1, we only need the average TP length to compute the throughput for the softmax policy. Suppose that the attacker uses its optimal strategy. An analysis similar to what we did in section 2, results in the following theorem.

Theorem 2: For $N = 2$, the average TP average length for a softmax policy with main probability q is given by:

$$L^s(q,d) = qL^m(\alpha d) + (1-q)L^n(\alpha(1-d)),$$

where $L^m(\cdot)$ is the function defined in (6) and $L^n(x) = 1 + \frac{p_{01}(1-x)}{1-p_{11}(1-x)}$.

Proof: Using a procedure similar to the one used in the proof of Theorem 1, we can readily show that if the channel selection algorithm always selects the channel with the smaller belief

value (which is the opposite of a greedy action), then the stationary distribution of $\{L_k\}_{k=1}^{\infty}$ is

$$\lambda_k = \begin{cases} 1 - p_{01}(1-\alpha), & k = 1 \\ p_{01}(1-\alpha)(1-p_{11}(1-\alpha))(p_{11}(1-\alpha))^{k-2} & k > 1 \end{cases} \tag{15}$$

Using (15) to compute $L^n(\alpha) = \sum_{k=1}^{\infty} k\lambda_k$ results in

$$L^n(\alpha) = 1 + \frac{p_{01}(1-\alpha)}{1 - p_{11}(1-\alpha)}.$$

By using the Bayes rule and the statement of Theorem 1, we can obtain the statement of Theorem 2. ■

To quantify how much randomness is added to the channel selection system by the softmax policy, we use the entropy of the channel selection probability distribution, $\mathcal{H}$. For the $N = 2$ case, because we use Bernoulli distribution, $\mathcal{H} = -(q\ln(q) + (1-q)\ln(1-q))$. By changing q from 1 to 0.5, the entropy increases from 0 to its maximum value $\ln(2)$.

We denote $U^{\pi}(\alpha)$ as the throughput of the cognitive radio, when it uses policy π for channel selection and the attacker uses its α-optimal strategy. We use the following definitions to quantify the robustness and the performance of policy π:

Definition 2: The **robustness** of a policy π for a channel selection system under an α-optimal attack is: $R^{\pi}(\alpha) = 1 - \frac{U^{\pi}(0) - U^{\pi}(\alpha)}{U^{\pi}(0)}$.

Definition 3: The **performance** of a policy π for a channel selection system under an α-optimal attack is: $P^{\pi} = U^{\pi}(0)$.

C. Numerical Results for two Channels

Using the results of Theorem 2, it can easily be shown that for $\alpha = 0$, $q^* = 1$, i.e., a non-adversarial environment the optimal softmax policy is equivalent to the myopic policy. Solving the optimization problems 1 and 2 for other values of α is straightforward due to the problem's small solution space. Figure 3 shows the solutions to this problem for a range of attack probabilities. Because myopic sensing is a special case of softmax sensing, it is obvious that $U^s \geq U^m$. Numerical results illustrated in Figure 3 shows that except for small values of α ($\alpha < 0.1$), U^s is strictly greater than U^m, i.e., softmax sensing outperforms myopic sensing.

Figure 4 shows the trade-off between the robustness and the performance of the system for fixed attack probability $\alpha = 0.5$. As it can be seen, increasing the randomness that is used in

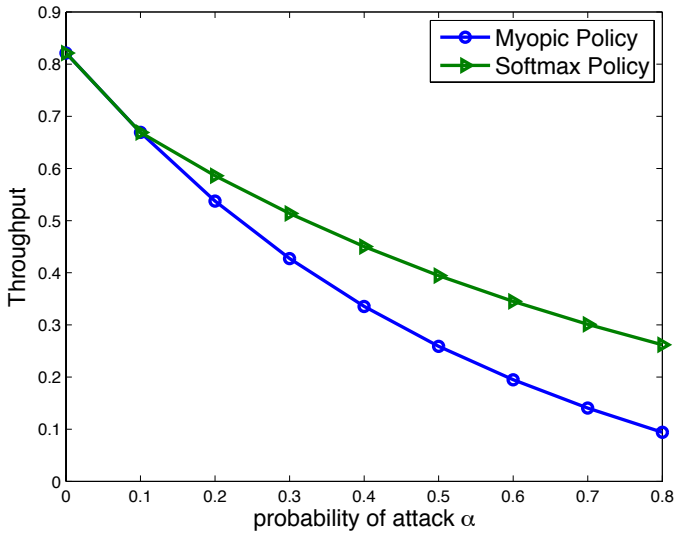


Fig. 3. Throughput vs. attack probability for $N = 2$.

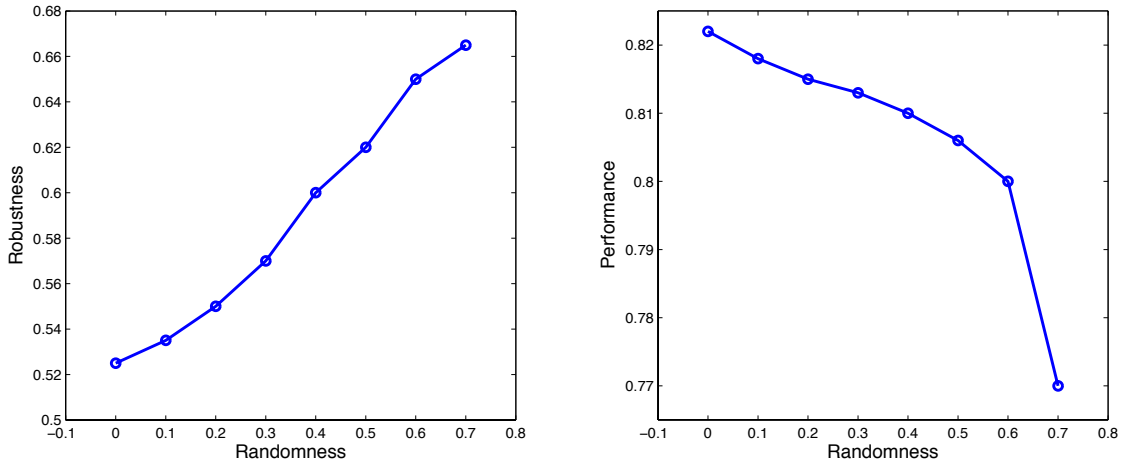


Fig. 4. Performance and robustness vs. randomness for $N = 2$.

the system by the softmax policy, increases the policy's robustness but it would decrease the policy's performance when no attacker exists.

V. Sensing Policies with More than Two Channels in an Adversarial Environment

As mentioned in [2], because L_k is a random process with higher-order memory, obtaining a closed-form expression of throughput for $N > 2$ channels with different statistics is very difficult. Nevertheless we can show that a softmax policy that uses a Boltzmann distribution with an intelligent choice for the temperature, τ, can outperform myopic policy for all possible strategies of the attacker, including its optimal strategy.

Assume that the channel selection system selects one of the N channels at each time slot. When a channel is selected, the system keeps using that channel until a primary user begins to transmit on the channel. Suppose $\Omega(t) = (\omega_1(t), \omega_2(t), \cdots, \omega_N(t))$ is the belief vector of the system at time t and $Q(t) = (q_1(t), q_2(t), \cdots, q_N(t))$ is the corresponding probability vector, i.e. $q_i(t) = f_i(\Omega(t))$ is the probability of selecting channel i at time t where $f(\cdot)$ is a function that maps the belief vector into a probability value. For the sake of notation simplicity, we omit t in the belief and and probability vectors henceforth. Without loss of generality, assume that $\omega_N \geq \omega_{N-1} \geq \cdots \geq \omega_1$ and consequently $q_N \geq q_{N-1} \geq \cdots \geq q_1$. As we mentioned in Section 2, the α-optimal strategy for the attacker is a strategy that minimizes the throughput while keeping the attack probability α fixed. In order to do so, the attacker needs to attack channel i with probability αd_i, where d_i is the division probability for channel i (i.e., the conditional probability of channel i being attacked at a given timeslot, assuming that the timeslot is attacked) which is a function of Q and Ω. Also to keep α fixed, we need to have $\sum_{i=1}^{N} d_i = 1$. The following theorem states the optimization problem that the attacker needs to solve to find its optimal strategy.

Theorem 3: In order to find its α-optimal strategy, the attacker needs to solve the following equality-constrained convex optimization problem:

Optimization Problem 3.

$$\min_{d_1,\cdots,d_N} \sum_{i=1}^{N} q_i L(\omega_i, \alpha d_i)$$

$$\text{s.t.} \sum_{i=1}^{N} d_i = 1, \forall i : d_i \geq 0$$

where

$$L(\omega_i, \alpha) = 1 + \frac{\omega_i(1-\alpha)}{1 - p_{11}^i(1-\alpha)}. \tag{16}$$

Proof: In order to minimize the throughput, attacker needs to minimize the expected value for $L_k(\omega)$. We know that the attacker attacks this channel with probability αd_i, so we have:

$$Pr\{L_k(\omega_i) = l\} = \begin{cases} 1 - \omega_i(1 - \alpha d_i), & l = 1 \\ \omega_i(1 - \alpha d_i)(p_{11}^i(1 - \alpha d_i))^{l-2} p_{10}^i, & l > 1 \end{cases}$$

It can easily be shown that

$$E[L_k(\omega)|\text{channel } i] = 1 + \frac{\omega_i(1 - \alpha d_i)}{1 - p_{11}^i(1 - \alpha d_i)}.$$

So we have:

$$\begin{aligned} E[L_k(\omega)] &= E[E[L_k(\omega)|\text{channel } i]] \\ &= \sum_{i=1}^{N} q_i(1 + \frac{\omega_i(1 - \alpha d_i)}{1 - p_{11}^i(1 - \alpha d_i)}). \end{aligned}$$

To prove that the $E[L_k(\omega)]$ is a convex function of d_i's, we compute the Hessian matrix of this function: $H_{N\times N} = [h_{ij}]$. We can see that the Hessian of this function is a diagonal matrix where:

$$h_{ii} = \frac{2q_i\omega_i\alpha^2 p_{11}^i}{(1 - p_{11}^i(1 - \alpha d_i))^3}.$$

It is obvious that we have $\forall i : h_{ii} \geq 0$, thus the matrix H is positive semidefinite and as a result $E[L_k(\omega)]$ is a convex function. ■

The above equality-constrained convex optimization problem can be solved by using elimination and Newton's method in $\frac{(N+1)^3}{3}$ steps [16]. However, solving this optimization problem requires attacker's exact knowledge about the channel selection strategy of the system and the statistics of all channels during a long period of time. Obviously, acquiring such knowledge is not feasible under most circumstances. Therefore instead of using the α-optimal strategy, the attacker can use alternative, simpler strategies. Each of these strategies differ in terms of the amount of knowledge on the channel selection system that is required. These strategies are:

- Greedy Strategy: The attacker only knows the best channel for transmission at each time slot and attacks this channel, i.e., $d_N = 1$ and $d_i = 0$ for $1 \leq i \leq N - 1$. This strategy is the optimal attack strategy when the channel selection system uses a myopic policy.
- Uniform Strategy: The adversary attacks all channels equiprobably, i.e., $d_i = \frac{1}{N} : 1 \leq i \leq N$.

This strategy can be used when the attacker does not have any knowledge about the channel selection system.

- Ω Strategy: The attacker only knows the channel statistics $\Omega = (\omega_1, \cdots, \omega_N)$ and has no knowledge about the channel selection policy of the system. It has a Boltzmann distribution with an arbitrary temperature τ_a, i.e., $d_i = \frac{e^{\omega_i/\tau_a}}{\sum_{j=1}^{N} e^{\omega_j/\tau_a}} : 1 \leq i \leq N$. Simulation results show that when the attack probability, α, is large, this strategy inflicts approximately the same effect as the α-optimal strategy.

In order to find the best channel selection strategy, the system assumes the worst-case attack scenario when the attacker uses its α-optimal strategy, which is the solution to the Optimization Problem 3: $d_i^* = f_i^*(Q, \Omega)$. The channel selection system needs to solve the following optimization problem in order to maximize the cognitive radio's throughput.

Optimization Problem 4.

$$\max_{q_1,\cdots,q_N} \sum_{i=1}^{N} q_i L(\omega_i, \alpha f_i^*(Q, \Omega))$$

$$\text{s.t.} \sum_{i=1}^{N} q_i = 1, \forall i : q_i \geq 0$$

Finding the global optimal solution of the non-linear optimization problem 4, gives us the amount of randomness that we need to add to the system in order to minimize the attack's effect and maximize the throughput.

We can also show that *for all possible attack strategies, including the α-optimal strategy, a softmax policy that uses a Boltzmann distribution with a well-chosen temperature outperforms myopic policy*.

Theorem 4: When the channel selection system is consisted of more than two identical channels, for all attack strategies that the adversary may employ, with a fixed attack probability $\alpha > \frac{(\omega_0 - p_{01})N}{(\omega_0 N - p_{01})}$, a softmax policy that uses a Boltzmann distribution with temperature

$$\tau > \frac{\omega_0 - p_{01}}{\ln \frac{p_{01}(N-\alpha)}{\omega_0 N(1-\alpha)}}, \tag{17}$$

achieves a greater throughput than a myopic policy.

Proof: Let $\omega_N \geq \omega_{N-1} \geq \cdots \geq \omega_1$ denote the belief values of all channels in the first slot

of the k-th TP. For the myopic policy, the length L_k of this TP has the following distribution.

$$Pr\{L_k(\omega_N) = l\} = \begin{cases} 1 - \omega_N(1-\alpha), & l = 1 \\ \omega_N(1-\alpha)(p_{11}(1-\alpha))^{l-2}p_{10}, & l > 1 \end{cases}$$

And for the softmax policy, the length L_k of this TP has the following distribution.

$$Pr\{L_k(\overline{\omega}) = l\} = \begin{cases} 1 - \overline{\omega}, & l = 1 \\ \overline{\omega}(p_{11}(1-\alpha\overline{d}))^{l-2}p_{10}, & l > 1 \end{cases},$$

where $\overline{\omega} = \sum_{i=1}^{N} q_i\omega_i(1-\alpha d_i)$ and $\overline{d} = \sum_{i=1}^{N} q_i d_i$ in which $q_i = \frac{e^{\omega_i/\tau}}{\sum_{j=1}^{N} e^{\omega_j/\tau}} = \frac{1}{\sum_{j=1}^{N} e^{(\omega_j-\omega_i)/\tau}}$ are the action selection probabilities that follow a Boltzmann distribution. It is readily observable that if $\overline{\omega} \geq \omega_N(1-\alpha)$, then $L_k(\overline{\omega})$ stochastically dominates $L_k(\omega_N)$ and consequently the throughput of the softmax policy would be greater than the throughput of the myopic policy. We use the fact that $\forall i$, $p_{01} \leq \omega_i \leq \omega_0$ which results in

$$\frac{1}{Ne^{(\omega_0-p_{01})/\tau}} \leq q_i \leq \frac{1}{Ne^{(p_{01}-\omega_0)/\tau}}, \tag{18}$$

and also the fact that $q_N > \frac{1}{N}$ and $d_N > \frac{1}{N}$.

Now we show that $\overline{\omega} \geq \omega_N(1-\alpha)$. Using (18) and (17), we have:

$$\begin{aligned}
\overline{\omega} &= \omega_N q_N(1-\alpha d_N) + \sum_{i=1}^{N-1} q_i\omega_i(1-\alpha d_i) \\
&\geq \omega_N q_N(1-\alpha d_N) + \sum_{i=1}^{N-1} p_{01} q_i(1-\alpha d_i) \\
&\geq \omega_N q_N(1-\alpha d_N) + \sum_{i=1}^{N-1} p_{01}\frac{1}{Ne^{(\omega_0-p_{01})/\tau}}(1-\alpha d_i) \\
&\geq \omega_N q_N(1-\alpha d_N) + \frac{p_{01}}{N}\frac{\omega_0 N(1-\alpha)}{p_{01}(N-\alpha)}(N-1-\alpha(1-d_N)) \\
&\geq \omega_N(q_N(1-\alpha d_N) + \frac{(1-\alpha)}{(N-\alpha)}(N-1-\alpha(1-d_N))) \\
&\geq \omega_N(\frac{1}{N}(1-\alpha) + \frac{(1-\alpha)}{(N-\alpha)}(N-1-\alpha(1-\frac{1}{N}))) \\
&= \omega_N(1-\alpha)
\end{aligned}$$

.

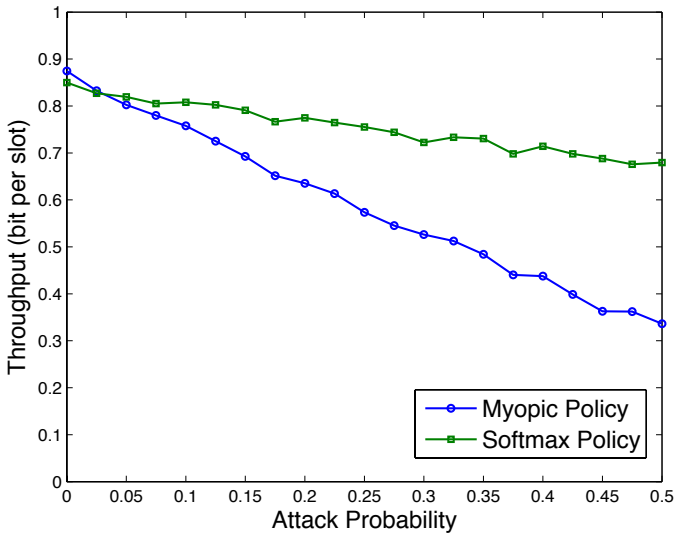


Fig. 5. Throughput vs. attack probability for N=4.

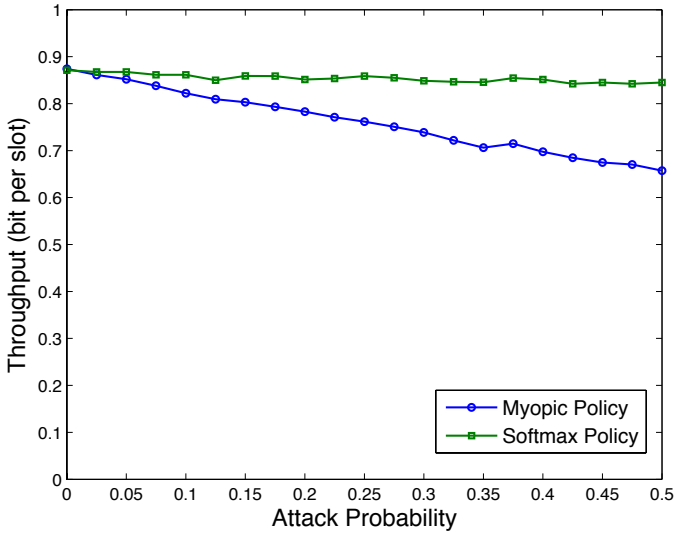


Fig. 6. Throughput vs. attack probability for N=10.

■

A. Numerical Results for More than Two Channels

We simulated the channel selection system to evaluate the performance of myopic sensing and softmax sensing for more than two channels. Figures 5 and 6 show the throughput of the channel

TABLE I
TRANSITION PROBABILITIES

	p_{11}	p_{10}	p_{00}	p_{01}
Channel 1	0.9	0.1	0.8	0.2
Channel 2	0.95	0.05	0.8	0.2
Channel 3	0.9	0.1	0.85	0.15
Channel 4	0.95	0.05	0.85	0.15

selection system versus the attack probability for myopic and softmax policies when 4 channels and 10 channels are available, respectively. To observe the effect of the number of channels on our scheme, we used identical channels with transition probabilities $p_{11} = 0.9$, $p_{10} = 0.1$, $p_{00} = 0.8$ and $p_{01} = 0.2$ to obtain the results in Figures 5 and 6. Softmax policies in these figures use a Boltzmann distribution with fixed temperature $\tau = 2$ for channel selection.

As it can be seen, the throughput drop caused by the increase in the attack probability is more severe for the myopic policy. Comparing the slope of lines in Figures 5 and 6, we can see that increasing the number of channels from 4 to 10 makes the softmax policy more robust against the belief manipulation attacks. In Figure 6, the softmax policy's drop in throughput is barely noticeable even as the attack probability is increased to 0.5.

Figures 5 and 6 also show that in a non-adversarial environment (i.e. when $\alpha = 0$), the performance of myopic policy is better than the performance of a softmax policy. These figures confirm the findings of [2] in which the authors showed that the optimal policy in non-adversarial environments is the myopic policy.

To compare different attack strategies and to investigate the amount of required randomness in the channel selection system, we used the more realistic model of non-identical channels with different transition probabilities. Figure 7 shows the effectiveness of the different attack strategies for different attack probabilities in a system of 4 non-identical channels with transition probabilities shown in Table 1.

As can be seen in Figure 7, the Ω strategys performance is close to that of the α-optimal strategy for large values of α. We can also see from the figure that the greedy strategy is the worst strategy among these four different strategies.

It is well known that the temperature of a Boltzmann distribution, τ, is a measure of the amount of randomness in the distribution, which can be inferred from the fact that the entropy

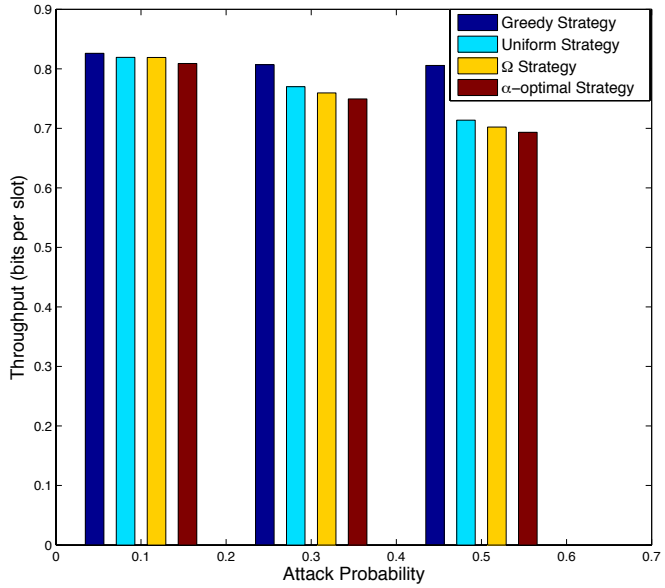


Fig. 7. Comparing different attack strategies.

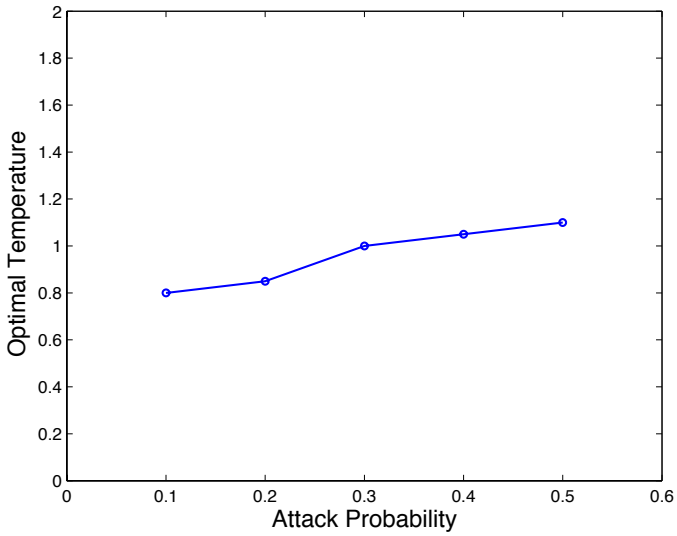


Fig. 8. Optimal amount of randomness vs. attack probability

of the Boltzmann distribution is an increasing function of τ [21]. Figure 8 plots the optimal temperature value for the channel selection system when the attacker uses the α-optimal strategy. These temperature values were obtained by solving Optimization Problem 4. From the figure,

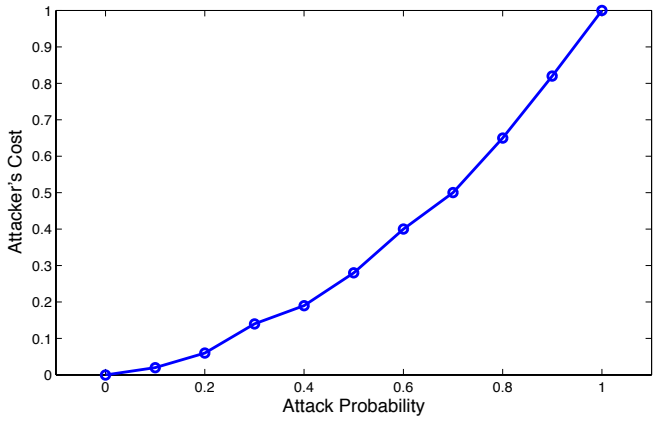


Fig. 9. Attacker's cost vs. attack probability

we can observe that the channel selection system needs to increase the randomness (i.e., τ) in its learning process as the attack probability (i.e., α) is increased to minimize the effect of the attack.

We observed the effect of increasing attack probability on performance of the channel selection system. Figure 9 plots the attacker's cost using the equation 4 versus the attack probability. This figure shows that though increasing the attack probability reduces the performance of the channel selection system significantly, it also increases the attacker's cost with a high rate.

VI. Conclusion

In this paper, we analyzed the security of a reinforcement learning algorithm that is used for solving the channel selection problem. We proposed a sensing policy that uses some level of randomness in the decision function to hide information about the learning algorithm. The obtained theoretical and simulation results show that the proposed mitigation technique can cause a dramatic improvement in the robustness of the channel selection process to adaptive jamming attacks. This countermeasure is applicable to other cognitive radio applications that use the same type of machine learning algorithm.

References

[1] Y. Chen, Q. Zhao, and A. Swami, "Joint design and separation principle for opportunistic spectrum access in the presence of sensing errors," IEEE Transactions on Information Theory, vol. 54, no. 5, May, 2008.

[2] Q. Zhao, B. Krishnamachari, and K. Liu, "On myopic sensing for multi- channel opportunistic access: structure, optimality, and performance," IEEE Transactions on Wireless Communications, 2008.

[3] S.H. Ahmad, M. Liu, T. Javidi, Q. Zhao and B. Krishnamachari, "Optimality of Myopic Sensing in Multi-Channel Opportunistic Access" , in IEEE Transactions on Information Theory, vol. 55, No. 9, pp. 4040-4050, September, 2009.

[4] S.H.A.Ahmad and M.Liu,"Multi-channel opportunistic access: a case of restless bandits with multiple plays," Allerton Conference, 2009.

[5] K. Liu and Q. Zhao, "A restless bandit formulation of opportunistic access: indexablity and index policy," Proc. of the 5th IEEE Conference on Sensor, Mesh and Ad Hoc Communications and Networks Workshops, June, 2008.

[6] L. Lai, H. ElGamal, H. Jiang, and V. Poor, "Cognitive Medium Access: Exploration, Exploitation and Competition," IEEE Transaction on Mobile Computing, 2011.

[7] D. Lowd and C. Meek. "Adversarial learning," In Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 641-647, 2005.

[8] M. Barreno, B. Nelson, R. Sears, A. Joseph, and J. Tygar. "Can machine learning be secure?," In ACM Symposium on Information, Computer and Communication Security, pages 16-25, 2006.

[9] B. Nelson, M. Barreno, F. J. Chi, A. D. Joseph, B. I. Rubinstein, U. Saini, C. Sutton, J. Tygar, and K. Xia. "Exploiting machine learning to subvert your spam filter," In Proceedings of the First USENIX Workshop on Large- Scale Exploits and Emergent Threats, April 2008.

[10] M. Barreno, B. Nelson, A. D. Joseph, and D. Tygar, "The security of machine learning," Machine Learning Journal (MLJ) Special Issue on Machine Learning in Adversarial Environments, 2008.

[11] K. Thomas, C. Grier, J. Ma, V. Paxson, and D. Song, "Design and Evaluation of a Real-Time URL Spam Filtering Service," IEEE Symposium on Security and Privacy 2011.

[12] E.N. Gilbert, "Capacity of burst-noise channels," Bell Syst. Tech. J., vol. 39, pp. 1253-1265, Sept. 1960.

[13] R. Smallwood and E. Sondik, "The optimal control of partially observable Markov processes over a finite horizon," Operations Research, pp. 1071-1088, 1971.

[14] R. S. Sutton and A. G. Bareto, "Reinforcement Learning: An Introduction," MIT press, 1998.

[15] A. Wald, "Sequential Analysis." Wiley, 1947.

[16] S. Boyd and L. Vandenberghe, "Convex Optimization," Cambridge University Press, 2004.

[17] N. Baldo and M. Zorzi, "Learning and adaptation in cognitive radios using neural networks," In: 1st IEEE workshop on cognitive radio networks (in conjunction with IEEE CCNC 2008), 12 January 2008, Las Vegas, Nevada, USA.

[18] T. Newman and T. Clancy, "Security threats to cognitive radio signal classifiers," in Virginia Tech Wireless Personal Communications Symposium, June 2009.

[19] T. Clancy, A. Khawar, and T. Newman, "Robust Signal Classification Using Unsupervised Learning," IEEE Transaction on Wireless Communication, April 2011.

[20] M. Li, I. Koutsopoulos, R. Poovendran, "Optimal Jamming Attack Strategies and Network Defense Policies in Wireless Sensor Networks," IEEE Transactions on Mobile Computing, vol. 9, no. 8, pp. 1119-1133, Apr. 2010

[21] M. Ohya, and M. Mizu, "An information-theoretical approach and its application to optimal problems," Electronics and Communication in Japan, Part 3, Vol 76, No. 6, 1993.